INS A1

A method of estimating the pitch of a speech signal using an average distance between peaks, use of the method, and a device adapted therefor

The invention relates to a method of estimating the pitch of a speech signal, said method being of the type where the speech signal is divided into segments, a conformity function for the signal is calculated for each segment, and peaks in the conformity function are detected. The invention also relates to the use of the method in a mobile telephone. Further, the invention relates to a device adapted to estimate the pitch of a speech signal.

In many speech processing systems it is desirable to know the pitch period of the speech. As an example, several speech enhancement algorithms are dependent on having a correct estimate of the pitch period. One field of application where speech processing algorithms are widely used is in mobile telephones.

A well known way of estimating the pitch period is to use the autocorrelation function, or a similar conformity function, on the speech signal. An example of such a method is described in the article D. A. Krubsack, R. J. Niederjohn, "An Autocorrelation Pitch Detector and Voicing Decision with Confidence Measures Developed for Noise-Corrupted Speech", IEEE Transactions on Signal Processing, vol. 39, no. 2, pp. 319-329, Febr. 1991. The speech signal is divided into segments of 51.2 ms, and the standard short-time autocorrelation function is calculated for each successive speech segment. A peak picking algorithm is applied to the autocorrelation function of each segment. This algorithm starts by choosing the maximum peak (largest value) in the pitch range of 50 to

333 Hz. The period corresponding to this peak is selected as an estimate of the pitch period.

However, such a basic pitch estimation algorithm is not sufficient. In some cases pitch doubling or pitch halving can occur, i.e. the highest peak appears at either half the pitch period or twice the pitch period. The highest peak may also appear at another multiple of the true pitch period. In these cases a simple selection of the maximum peak will provide a wrong estimate of the pitch period.

The above-mentioned article also discloses a method of improving the algorithm in these situations. The algorithm checks for peaks at one-half, one-third, one-fourth, one-fifth, and one-sixth of the first estimate of the pitch period. If the half of the first estimate is within the pitch range, the maximum value of the autocorrelation within an interval around this half value is located. If this new peak is greater than one-half of the old peak, the new corresponding value replaces the old estimate, thus providing a new estimate which is presumably corrected for the possibility of the pitch period doubling error. This test is performed again to check for double doubling errors (fourfold errors). If this most recent test fails, a similar test is performed for tripling errors of this new estimate. This test checks for pitch period errors of sixfold. If the original test failed, the original estimate is tested (in a similar manner) for tripling errors and errors of fivefold. The final value is used to calculate the pitch estimate.

However, this known algorithm is rather complex and requires a high number of calculations, and these drawbacks make it less usable in real time environments on small digital signal processors as they are used in mobile

telephones and similar devices. Further, the algorithm only checks for pitch doubling, pitch tripling, etc., while pitch halving is not considered. Actually, if a peak is present at the half of the true pitch period, the algorithm would (wrongly) choose that peak as the estimate of the pitch period.

Thus, it is an object of the invention to provide a method of the above-mentioned type which is less complex than the prior art methods, such that the method is suitable for small digital signal processors. Further, the method should also avoid the pitch halving situation.

According to the invention, this object is achieved in that the method comprises the steps of estimating an average distance between the detected peaks, and using this estimate of the average distance as an estimate of the pitch.

When the method is based on the fact that the identified peaks in the conformity function show a periodic behaviour and that the true pitch period actually corresponds to the distance between the peaks, a simpler algorithm is achieved which provides the true pitch period independent on the occurrence of pitch halving, pitch doubling, etc.

When the method further comprises the steps of sampling the speech signal to obtain a series of samples, and performing the division into segments such that each segment has a fixed number of consecutive samples, an even less complex method is achieved because only a finite number of samples has to be considered.

When the method further comprises the steps of estimating a set of filter parameters using linear predictive analysis (LPA), providing a modified signal by filtering the

speech signal through a filter based on this estimated set of filter parameters, and calculating the conformity function of the modified signal, much of the smearing of the original speech signal is removed and thus the possibility of clearer peaks in the conformity function is improved, which results in a more precise estimation of the pitch period.

An expedient embodiment of the invention is achieved when the conformity function is calculated as an autocorrelation function. However, it should be noted that also other conformity functions may be utilized, such as e.g. a cross correlation between the original speech signal and the above-mentioned modified signal.

An improved method is achieved when the method further comprises the steps of calculating, for each peak in the conformity function, the difference between the position of the peak and the estimate of the average distance, and providing an improved estimate of the pitch by selecting as the improved estimate the position of the peak having the smallest value of said difference. In this way the position of an actual peak is used as the estimate and it is still assured that the correct peak is used. If, in this case, the peak having the smallest value of the difference is represented by a number of samples, the best estimate is achieved when the sample having the maximum amplitude of the conformity function is selected as the improved estimate of the pitch.

In an expedient embodiment of the invention the method is used in a mobile telephone which is a typical example of a device having only limited computational resources.

As mentioned, the invention further relates to a device adapted to estimate the pitch of a speech signal. The de-

vice comprises means for dividing the speech signal into segments, means for calculating for each segment a conformity function for the signal, and means for detecting peaks in the conformity function. When the device is further adapted to estimate an average distance between said peaks, and to use the estimate of said average distance as an estimate of the pitch, a device less complex than prior art devices is achieved, which also avoids the pitch halving situation.

When the device further comprises means for sampling the speech signal to obtain a series of samples, and means for performing said division into segments such that each segment has a fixed number of consecutive samples, an even less complex device is achieved because only a finite number of samples has to be considered.

When the device further comprises means for estimating a set of filter parameters using linear predictive analysis (LPA), means for providing a modified signal by filtering the speech signal through a filter based on this estimated set of filter parameters, and means for calculating the conformity function of the modified signal, much of the smearing of the original speech signal is removed and thus the possibility of clearer peaks in the conformity function is improved, which results in a more precise estimation of the pitch period.

An expedient embodiment of the invention is achieved when the conformity function is an autocorrelation function. However, it should be noted that also other conformity functions may be utilized, such as e.g. a cross correlation between the original speech signal and the above-mentioned modified signal.

An improved device is achieved when the device further comprises means for calculating, for each peak in the conformity function, the difference between the position of the peak and the estimate of the average distance, and means for providing an improved estimate of the pitch by selecting as the improved estimate the position of the peak having the smallest value of said difference. In this way the position of an actual peak is used as the estimate and it is still assured that the correct peak is used. If, in this case, the peak having the smallest value of the difference is represented by a number of samples, the best estimate is achieved when the sample having the maximum amplitude of the conformity function is selected as the improved estimate of the pitch.

In an expedient embodiment of the invention, the device is a mobile telephone, which is a typical example of a device having only limited computational resources.

In another embodiment the device is an integrated circuit which can be used in different types of equipment.

The invention will now be described more fully below with reference to the drawing, in which

figure 1 shows a block diagram of a pitch detector according to the invention,

figure 2 shows the generation of a residual signal,

figure 3a shows a 20 ms segment of a voiced speech signal,

figure 3b shows the autocorrelation function of a residual signal corresponding to the segment of figure 3a,

figure 4 shows an example of an autocorrelation function where pitch doubling could arise, and

figure 5 shows an example of the calculation of the distance between peaks in an autocorrelation function.

Figure 1 shows a block diagram of an example of a pitch detector 1 according to the invention. A speech signal 2 is sampled with a sampling rate of 8 kHz in the sampling circuit 3 and the samples are divided into segments or frames of 160 consecutive samples. Thus, each segment corresponds to 20 ms of the speech signal. This is the sampling and segmentation normally used for the speech processing in a standard mobile telephone.

Each segment of 160 samples is then processed in a filter 4, which will be described in further detail below.

First, however, the nature of speech signals will be mentioned briefly. In a classical approach a speech signal is modelled as an output of a slowly time-varying linear filter. The filter is either excited by a quasi-periodic sequence of pulses or random noise depending on whether a voiced or an unvoiced sound is to be created. The pulse train which creates voiced sounds is produced by pressing air out of the lungs through the vibrating vocal cords. The period of time between the pulses is called the pitch period and is of great importance for the singularity of the speech. On the other hand, unvoiced sounds are generated by forming a constriction in the vocal tract and produce turbulence by forcing air through the constriction at a high velocity. This description deals with the detection of the pitch period of voiced sounds and thus, unvoiced sounds will not be further considered.

As speech is a varying signal also the filter has to be time-varying. However, the properties of a speech signal change relatively slowly with time. It is reasonable to believe that the general properties of speech remain fixed for periods of 10-20 ms. This has led to the basic principle that if short segments of the speech signal are considered, each segment can effectively be modelled as having been generated by exciting a linear time-invariant system during that period of time. The effect of the filter can be seen as caused by the vocal tract, the tongue, the mouth and the lips.

As mentioned, voiced speech can be interpreted as the output signal from a linear filter driven by an excitation signal. This is shown in the upper part of figure 2 in which the pulse train 21 is processed by the filter 22 to produce the voiced speech signal 23. A good signal for the detection of the pitch period is obtained if the excitation signal can be extracted from the speech. By estimating the filter parameters A in the block 24 and then filtering the speech through an inverse filter 25 based on the estimated filter parameters, a signal 26 similar to the excitation signal can be obtained. This signal is called the residual signal. This process is shown in the lower part of figure 2. The blocks 24 and 25 are included in the filter 4 in figure 1.

The estimation of the filter parameters is based on an all-pole modelling which is performed by means of the method called linear predictive analysis (LPA). The name comes from the fact that the method is equivalent with linear prediction. This method is well known in the art and will not be described in further detail here.

The estimation of the pitch is based on the autocorrelation of the residual signal, which is obtained as de-

scribed above. Thus, the output signal from the filter 4 is taken to an autocorrelation calculation unit 5. Figure 3a shows an example of a 20 ms segment of a voiced speech signal and figure 3b the corresponding autocorrelation function of the residual signal. It will be seen from figure 3a that the actual pitch period is about 5.25 ms corresponding to 42 samples, and thus the pitch estimation should end up with this value.

The next step in the estimation of the pitch is to apply a peak picking algorithm to the autocorrelation function provided by the unit 5. This is done in the peak detector 6 which identifies the maximum peak (i.e. the largest value) in the autocorrelation function. The index value, i.e. the sample number or the lag, of the maximum peak is then used as a preliminary estimate of the pitch period. In the case shown in figure 3b it will be seen that the maximum peak is actually located at a lag of 42 samples. The search of the maximum peak is only performed in the range where a pitch period is likely to be located. In this case the range is set to 60-333 Hz.

However, this basic pitch estimation algorithm is not always sufficient. In some cases pitch doubling or halving may occur, i.e. due to distortion the peak in the autocorrelation function corresponding to the true pitch period is not the highest peak, but instead the highest peak appears at either half the pitch period or twice the pitch period. The highest peak could also appear at other multiples of the actual pitch period (pitch tripling, etc.) although this occurs relatively rarely. A typical example where pitch doubling would arise is shown in figure 4 which again shows the autocorrelation function of the residual signal. Here too, the correct pitch period would be around 42 samples, but the peak at twice the pitch period, i.e. around 84 samples, is actually higher

than the one at 42 samples. The basic pitch estimation algorithm would therefore estimate the pitch period to 84 samples and pitch doubling would thus occur. It will also be seen that two smaller peaks are located around half the pitch period, and in some cases one of these could be higher than the correct peak and pitch halving would occur.

To avoid the problem of pitch doubling and halving the pitch detection algorithm is therefore improved as described below.

After the preliminary pitch estimate has been determined, it is checked in the risk check unit 7 whether there is any risk of pitch halving or pitch doubling. All peaks with a peak value higher than 75% of the maximum peak are detected and the further processing depends on the result of this detection. If only one peak is detected, i.e. the original maximum peak, there is no need to perform a process to avoid pitch doubling and pitch halving. In this situation the preliminary pitch estimate is used as the final pitch estimate. If, however, more than one peak is detected, there is a risk of pitch doubling or pitch halving, and a further algorithm must be performed to ensure that the correct peak is selected as the pitch estimate.

Two different solutions to such an algorithm will be described. One solution, which is performed in the unit 8, is used when pitch estimates are available from a number of previous segments, while the other solution, which is performed in the unit 9, is used when such estimates are not available, which will be the case in the beginning of a speech signal. The latter solution is described first.

In cases where no previously estimated pitch periods are available, the procedure to avoid pitch doubling and pitch halving is based on the fact that the identified peaks show a periodic behaviour. Actually it can be said that the pitch period simply corresponds to the distance between the peaks. Index values, i.e. the lag, of the detected peaks are sorted into groups depending on how close to each other the indexes are. In many cases a peak can be represented by more than one index, i.e. more than one sample, resulting in several indexes around a peak being detected. Indexes with a distance of less than e.g. five samples are sorted into the same group.

For each group an average is calculated and then differences (distances) between the averaged indexes are calculated. The difference towards zero is also calculated since the first peak may be the actual pitch period. If the detected peaks represent the periodic behaviour of the speech signal in the current segment the differences between the groups ought to be about the same.

Therefore, if the variance of the differences between the groups is below a given threshold, e.g. 10, the average of the differences, i.e. the average distance, is assumed to be approximately the pitch period and is thus used as a secondary estimate of the pitch period. The variance threshold can be set from watching probable differences between mean values and their variance.

An example of this procedure is shown in figure 5 in which level I shows the received indexes of the highest peaks. In level II the indexes are sorted into groups and the mean values of the groups are calculated in level III. The differences between mean values are shown in level IV and finally, the variance is calculated in level V.

The average distance may be used directly as the pitch estimate, or the method can be improved by subtracting the average distance from each of the average indexes representing different groups (level III). The group in which the smallest result of this subtraction, i.e. the group closest to the average distance, is found is selected as the pitch estimate.

If, however, the variance is above the threshold, it means that the distances between peaks are too different to represent the periodic behaviour of the signal. In this case the method cannot be used and the preliminary pitch estimate is maintained as the best estimate.

When this method has been used for a number of consecutive segments, and if the pitch estimates for these segments are stored in a memory, these previous estimates may be used in a different method of avoiding pitch doubling and pitch halving. This method is described below.

First, an average of the previous pitch estimates from e.g. the last 15 segments is calculated. This value is then subtracted from the index values where the highest peaks in the autocorrelation function of the residual signal are located, which means that the differences between the index values of the highest peaks and the average of the previously detected pitch periods are calculated. Since the pitch period for a given person is relatively constant over time, a small difference between the correct pitch period of the current segment and the average of the previous pitch estimates is expected. Therefore, those values in the resulting vector of subtraction results that are below a given threshold, e.g. 10, are selected. The use of the threshold is due to the fact that the pitch period may actually vary slightly while a

person is talking, and therefore such a difference has to be accepted. The actual threshold can be set from watching probable examples.

If only one difference is below the threshold the corresponding index value or lag is selected as the estimate of the pitch period. If more than one difference is below the threshold, the one with the highest amplitude in the autocorrelation of the residual signal is selected. If there are no differences below the threshold, this indicates that the pitch has changed drastically, as it may e.g. be the case when switching speakers. In such a case the preliminary pitch estimate is maintained as the best estimate.

This method utilizing previous estimates is considerably less complex than the other one based on the distance between the peaks, and therefore it should be used as soon as there are sufficient previous estimates in order to reduce the needed amount of computational resources.

As mentioned above, one example of equipment in which the invention can be implemented is a mobile telephone. The algorithm may also be implemented in an integrated circuit which may then be used in other types of equipment.

Although a preferred embodiment of the present invention has been described and shown, the invention is not restricted to it, but may also be embodied in other ways within the scope of the subject-matter defined in the following claims.

Thus, the autocorrelation function may be calculated directly of the speech signal instead of the residual signal, or other conformity functions may be used instead of the autocorrelation function. As an example, a cross cor-

relation could be calculated between the speech signal and the residual signal. It is also possible to repeat the autocorrelation, i.e. to calculate the autocorrelation of the result of the first autocorrelation, before detecting peaks.

Further, different sampling rates and sizes of the segments may be used.